Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Getter Põru

Veekogude klassifitseerimine satelliidiandmetelt multinomiaalse logistilise mudeliga

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendajad: Ene Käärik, PhD
Kristi Uudeberg, MSc

Tartu 2018

Veekogude klassifitseerimine satelliidiandmetelt multinomiaalse logistilise mudeliga

Bakalaureusetöö

Getter Põru

Lühikokkuvõte. Bakalaureusetöö eesmärk on luua statistiline mudel, mis satelliitsensori poolt mõõdetavate peegeldumisspektrite põhjal prognoosib iga optiliselt keerulise veekogu satelliitpildi piksli jaoks, millisesse veetüüpi ta kõige tõenäolisemalt kuulub. Töö teoreetilises osas antakse ülevaade veekogude kaugseirest ja tutvustatakse peakomponentide analüüsi ning multinomiaalset logistilist mudelit. Töö praktilises osas luuakse maapealsete mõõtmiste põhjal mudel, mida on võimalik rakendada satelliidiandmetele.

CERCS teaduseriala: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Märksõnad. Statistiline analüüs, kaugseire, pinnavesi, prognoosmudelid

Classifying Water Types from Satellite Data Using Multinomial Logistic Model

Bachelor's thesis

Getter Põru

Abstract. The aim of this thesis is to create a statistical model for the pixels in satellite images that predicts optically complex water bodies' most likely water type through the use of reflectance spectra measured by satellite sensors. The theoretical part of this thesis will give an overview of remote sensing. In addition, the principal component analysis and the multinomial logistic model will be introduced. The practical side of this thesis will involve creating a model that can be implemented on satellite information by using data from the surface of the earth

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics

Key words. Statistical analysis, remote sensing, surface water, predictive models

Sisukord

Sissejuhatus

Satelliitkaugseire kasutamine Maal toimuvate protsesside jälgimiseks on kiiresti kasvanud seoses kosmosetehnoloogia arenguga. Üks kaugseire valdkondi on veekogude dünaamika ja seisukorra uurimine. Esimene veekaugseireks sobiliku sensoriga varustatud satelliit saadeti orbiidile 1978. aastal NASA (*The National Aeronautics and Space Agency*) poolt. Aastal 2016 saadeti Euroopa Kosmoseagentuuri programmi Copernicus raames orbiidile peamiselt veekaugseireks mõeldud satelliitide seeria Sentinel-3 esimene satelliit, mille andmeid ka Eesti teadlased meie veekogude kaugseireks kasutada saavad. Käesolevas töös saadud mudeli loomisel on lähtutud satelliitsensorilt OLCI saadavate andmete eripärast.

Bakalaureusetöö eesmärk on luua mudel, mis satelliitsensori poolt mõõdetavate peegeldumisspektrite põhjal prognoosib iga optiliselt keerulise veekogu satelliitpildi piksli jaoks, millisesse veetüüpi ta kõige tõenäolisemalt kuulub. Veetüübi määramine optilisel viisil on oluline osa vee seisukorra hindamisest ja veetüübipõhiste algoritmide rakendamisest.

Töö esimeses peatükis antakse ülevaade veekogude optilistest omadustest ning satelliitkaugseirest. Teises peatükis tutvustatakse peakomponentide analüüsi ning multinomiaalse logistilise regressiooni metoodikat. Töö kolmandas peatükis kirjeldatakse peakomponentide analüüsi rakendamist ja multinomiaalse logistilise regressioonimudeli loomist ning analüüsitakse saadud tulemusi.

Töö on vormistatud kasutades tarkvaraprogrammi LaTeX ning andmete analüüsimiseks ja jooniste tegemiseks on kasutatud statistikatarkvara R.

Autor tänab Tartu Ülikooli Tartu Observatooriumi veekogude kaugseire töörühma andmete kogumise eest ning Eesti Teadusagentuuri (PSG10). Antud lõputöös on kasutatud andmeid, mis on kogutud Euroopa Liidu Horisont 2020 teadus- ja innovatsiooniprogrammi grandi nr 730066 ja Euroopa Liidu Regionaalarengu Fondi programmi „Keskkonnakaitse ja -tehnoloogia teadus ja arendustegevus“ (KESTA) projekti „Eesti veekeskkonna observatoorium“ (VeeOBS 3.2.0802.11-0043) raames.

Autor tänab ka töö juhendajaid Ene Käärikut ja Kristi Uudebergi kannatliku meele ning hea nõustamise eest.

1 Veekogude kaugseire

1.1 Satelliitkaugseire

Kaugseire oma olemuselt on valdkond, mis hõlmab andmete kogumist ja analüüsimist ilma uuritava objekti endaga füüsiliselt kokku puutumata. Kaugseire jaguneb kaheks: aktiivne ja passiivne. Aktiivse kaugseire puhul mõõdetakse kõrgemal platvormil või satelliidil paikneva sensori poolt spetsiaalselt tekitatud kiirgussignaali peegeldumist objektilt. Passiivse kaugseire korral mõõdab sensor uuritavalt objektilt peegeldunud päikesekiirgust või objektilt lähtuvat kiirgust. Veekogude kaugseires kasutatakse passiivse kaugseire optilist osa. Veekogudest lähtuv signaal on tunduvalt nõrgem kui maismaalt lähtuv signaal (lisa 1) ja lisaks ligi 90% satelliidil paikneva sensori poolt mõõdetud signaalist ei ole kontaktis veepinnaga, vaid on pärit atmosfäärist [1]. Seetõttu veekaugseireks sobilikel satelliitsensoritel on võrreldes maismaasensoritega teistsugune ruumiline, radiomeetriline ja spektraalne lahutus ning oluline on ka korrektne atmosfäärikorrektsioon.

Esimene veekogude uurimiseks mõeldud satelliitsensor CZCS (*Coastal Zone Color Scanner*) saadeti NASA poolt orbiidile 1978. aastal. Oma piiratud spektraalsele lahutusvõimele vaatamata, sensoril oli vaid kuus kanalit, oli see edukas veekvaliteedi omaduste määramise katsetustel ja meetodite edasiarendamisel [2]. Teaduse ja tehnika areng on olnud kiire ning sellest ajast alates on mitmed satelliidid varustatud veekogude uurimiseks sobilike sensoritega. Aastal 2016 saadeti Euroopa Kosmoseagentuuri (*European Space Agency*, ESA) programmi Copernicus raames orbiidile peamiselt veekaugseireks mõeldud satelliitide seeria Sentinel-3 esimene satelliit Sentinel-3A ning seeria teine satelliit Sentinel-3B saadeti orbiidile 2018. aastal. Mõlema satelliidi pardal on keskmise lahutusvõimega spektromeeter OLCI (*Ocean and Land Colour Instrument*). OLCI-l on 21 kanalit, mille lainepikkused varieeruvad spektri optilisest osast kuni lähisinfrapunakiirguseni ning kanalite laiused on vahemikus 2,5–40 nanomeetrit [3].

1.2 Vee optilised omadused

Veekogude kaugseire mõõtmistulemusi mõjutavad konkreetse veekogu optilised omadused (lisa 1). Loodusliku vee värvus ja läbipaistvus on määratud vee ja temas sisalduvate lahustunud ja lahustumata aine, orgaanilise ja mitteorgaanilise aine ning elus ja elutu aine neeldumise ja hajumise poolt. Neeldumis- ja hajumiskoefitsiente käsitletakse kui esmaseid optilisi omadusi (*inherent optical properties*, IOP), sest need sõltuvad ainult keskkonna keemilisest koostisest ja füüsikalistest omadustest. Vees leiduvaid osakesi, mis suudavad vette jõudnud kiirgust hajutada või neelata, nimetatakse optiliselt aktiivseteks aineteks (*optically active substances*, OAS). Rannikuvete ja järvede korral on neid kolm: fütoplankton (*phytoplankton*), värvunud lahustunud orgaaniline aine (*coloured dissolved organic matter*, CDOM) ja heljum (*total suspended matter*, TSM). Nende spektraalsed omadused määravad peamiselt optilise kaugseire põhilise uuritava parameetri, vee peegeldumisspektri, mis kuulub tuletatud optiliste omaduste (*apparent optical properties*, AOP) hulka, sest sõltub lisaks keskkonna enda optilistele omadustele ka valguse langemisest veepinnale.

Peegeldumisspekter R on defineeritud alt üles suunatud kiiritustiheduse E_u (*upwelling irradiance*) ja ülevalt alla suunatud kiiritustiheduse E_d (*downwelling irradiance*) suhtena sügavusel z lainepikkusel λ:

$$R(z,\lambda) = \frac{E_u(z,\lambda)}{E_d(z,\lambda)}.$$

Kaugseire peegeldumisspektrit R_{rs} (*remote-sensing reflectance*) kirjeldatakse veepinna kohal mõõdetud kirkuse L_w (*water leaving radiance*) ja veepinna kohal mõõdetud kogu ülevalt alla suunatud kiiritustiheduse E_d suhtena:

$$R_{rs}(\lambda) = \frac{L_w(z=0+,\lambda)}{E_d(z=0+,\lambda)},$$

kus parameeter $z = 0+$ tähistab vahetult veepinna kohal tehtud mõõtmisi. Lihtsustatud seose esmaste ja tuletatud optiliste omaduste vahel on esitanud Gordon jt (1975) [4]:

$$R(0-,\lambda) = \frac{E_u(z=0-,\lambda)}{E_d(z=0-,\lambda)} = C\frac{b_b}{a+b_b},$$

kus parameeter $z = 0-$ tähistab vahetult veepinna all tehtud mõõtmisi ning b_b on

tagasihajumiskoefitsient, a neeldumiskoefitsient ja C on koefitsient, mis sõltub kiirguse vette tungimise tingimustest.

1.3 Veekogude optiline klassifikatsioon

Vastavalt Morel & Prieur 1977. aastal avaldatud artiklile [5] on võimalik maailma veekogud optiliselt aktiivsete ainete sisalduse järgi jagada *Case 1* ja *Case 2* klassi. *Case 1* tüüpi veekogudes on vee optilised omadused määratud fütoplanktoni ja selle laguproduktide poolt ning sinna kuuluvad peamiselt ookeanid ning vaid mõned üksikud selgemad sise- ja rannikuveekogud. *Case 2* tüüpi veekogudes, kuhu kuulub enamik sise- ja rannikuveekogusid, mõjutavad vee optilisi omadusi lisaks klorofüllile ka värvunud lahustunud orgaaniline aine ja heljum.

Eesti ja selle lähiümbruskonna järved ja rannikuveed kuuluvad *Case 2* tüüpi veekogude hulka ja on optiliste omaduste poolest keerukad, sest optiliselt aktiivsete ainete kontsentratsioonid võivad üksteisest sõltumatult varieeruda veekogudes erinevalt. Seetõttu on vaja veel täpsustavamaid klassifikatsioone. Reinart jt [6] on näidanud, et optiliselt aktiivsete ainete kontsentratsiooonide põhjal on võimalik Lõuna-Soome ja Eesti järved jagada viide statistiliselt oluliselt erinevasse veetüüpi: *Selge* (*Clear*, C), *Mõõdukas* (*Moderate*, M), *Sogane* (*Turbid*, T), *Väga sogane* (*Very turbid*, V) ja *Pruun* (*Brown*, B). Samuti on kirjeldatud igale veetüübile omase peegeldumisspektri kuju. Seda aluseks võttes on Uudeberg jt [7] demonstreerinud võimalikkust veekogud jagada samasse viide klassi peegeldumisspektri põhjal. Antud liigitust on kasutatud ka käesolevas töös loodud mudeli koostamisel.

2 Kasutatav metoodika

2.1 Multinomiaalne logistiline regressioon

Binaarseks tunnuseks nimetatakse uuritavat tunnust Y, kui tal on kaks võimalikku väärtust. Tavaliselt kodeeritakse väärtused 1/0 selliselt, et 1 tähistab sündmuse toimumist ja 0 mitte toimumist ning sündmuse Y toimumise tõenäosus on $P(Y=1)=\pi$ ja mittetoimumise tõenäosus $P(Y=0)=1-\pi$. Sellisel juhul on tegemist binoomjaotusega $Y \sim B(n,\pi)$, kus n on katsete arv ning π sündmuse toimumise tõenäosus. Binaarse logistilise regressioonimudeliga hinnatakse sündmuse esinemise šansi $\frac{\pi}{1-\pi}$ logaritmi.

Uuritaval tunnusel võib tihti olla rohkem kui kaks taset. Sellisel juhul ei ole enam tegemist binoomjaotusega, vaid selle mitmemõõtmelise üldistuse multinoomjaotusega. Teiste sõnadega on multinoomjaotus binoomjaotuse üldistus. Analoogselt on multinomiaalne logistiline regressioon binaarse logistilise regressiooni üldistus, kus hinnatakse k taseme korral $k-1$ logistilist mudelit. Iga mudeli korral hinnatakse tasemele kuulumise šansi logaritmi mingi kindla baastaseme suhtes.

2.1.1 Multinoomjaotus

Käesolev alajaotus põhineb G. Tutzi raamatul „*Regression for Categorical Data*“ [8].

Multinoomjaotus on binoomjaotuse üldistus, kus igas üksikus katses on enam kui kaks võimalikku katsetulemust. Olgu võimalikud katsetulemused $1,\ldots,k$ ning nende esinemise tõenäosused vastavalt $\pi=(\pi_1,\ldots,\pi_k)$, kusjuures $\pi_1+\ldots+\pi_k=1$ ja $\forall i \in \{1,\ldots,k\} \quad \pi_i \in [0,1]$. Tõenäosus, et n sõltumatus katses sündmused $1,\ldots,k$ toimuvad vastavalt $\mathbf{y}^T=(y_1,\ldots,y_k)$ korda, avaldub valemiga

$$P_n(y_1,y_2,\ldots,y_k)=\frac{n!}{y_1!\ y_2!\ \ldots\ y_k!}\pi_1^{y_1}\ \pi_2^{y_2}\ \ldots\ \pi_k^{y_k}.$$

Kuna $\sum_{i=1}^{k} y_i = n$, siis on ühe sündmuse toimumiste arv avaldatav teiste kaudu ning kasutusele saab võtta lühema vektori $\mathbf{y}^T=(y_1,\ldots,y_{k-1})$. Siis tõenäosusfunktsioon avaldub

valemiga

$$P_n\left(y_1,\ldots,y_{k-1}\right)=\frac{n!}{y_1!\ \ldots\ y_{k-1}!\,(n-y_1-\ldots-y_{k-1})!}\pi_1^{y_1}\ \ldots\ \pi_{k-1}^{y_{k-1}}\cdot \\ \cdot\left(1-\pi_1-\ldots-\pi_{k-1}\right)^{n-y_1-\ldots-y_{k-1}}.$$

Multinoomjaotuse korral vektori $\mathbf{y}^T=(y_1,\ldots,y_{k-1})$ komponentide keskväärtus ja dispersioon avalduvad vastavalt $E(y_i)=n\pi_i$ ja $D(y_i)=n\pi_i(1-\pi_i)$ iga $i=1,\ldots,k-1$ korral.

2.1.2 Multinomiaalne *logit* mudel

Käesolev alajaotus on kirjutatud G. Tutzi raamatu „*Regression for Categorical Data*“ [8] ning S. A. Czepieli artikli „*Maximum likelihood estimation of logistic regression models: theory and implementation*“ [9] põhjal.

Olgu võimalikud katsetulemused taas $1,\ldots,k$ ning seletavate tunnuste arv m. Binaarse *logit* mudeli korral hinnatakse uuritava sündmuse toimumise šansi logaritmi

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right)=\beta_0+\beta_1x_1+\ldots+\beta_mx_m.$$

Multinomiaalse *logit* mudeli korral vaadeldakse $k-1$ *logit* mudelit, kus igas mudelis hinnatakse sündmuse toimumise ehk mingile kindlale tasemele kuulumise ja baastasemele kuulumise tõenäosuste suhte logaritmi.

Valides baastasemeks taseme k, avaldub r-ndale tasemele vastav *logit* mudel järgmiselt:

$$\ln\left(\frac{P(Y=r)}{P(Y=k)}\right)=\beta_{r0}+\beta_{r1}x_1+\ldots+\beta_{rm}x_m, \tag{1}$$

kus $r=1,\ldots,k-1$. Siinkohal tasub tähele panna, et parameetrid $\beta_{r0},\ldots,\beta_{rm}$ sõltuvad tasemest r ning baastaseme k võib valida vabalt tasemete $1,\ldots,k$ hulgast.

Tähistagu π_{ir} i-nda objekti tõenäosust kuuluda tasemele r. Valemist 1 saame siis

$$\ln\left(\frac{\pi_{ir}}{\pi_{ik}}\right)=\beta_{r0}+\beta_{r1}x_{i1}+\ldots+\beta_{rm}x_{im}.$$

Avaldades eelnevast π_{ir} ning võttes, et iga $i = 1, \ldots, n$ korral $x_{i0} = 1$, saame

$$\pi_{ir} = \frac{e^{\sum_{j=0}^{m} \beta_{rj} x_{ij}}}{1 + \sum_{r=1}^{k-1} e^{\sum_{j=0}^{m} \beta_{rj} x_{ij}}}, \tag{2}$$

kus $r = 1, \ldots, k-1$ ja

$$\pi_{ik} = 1 - \pi_{i1} - \ldots \pi_{i(k-1)} = \frac{1}{1 + \sum_{r=1}^{k-1} e^{\sum_{j=0}^{m} \beta_{rj} x_{ij}}}. \tag{3}$$

Valemid 2 ja 3 annavad prognoosi i-nda objekti tõenäosusele kuuluda vastavalt r-ndale või k-ndale uuritava tunnuse tasemele.

2.1.3 Parameetrite hindamine suurima tõepära meetodil

Käesolev alajaotus põhineb S. A. Czepieli artiklil „*Maximum likelihood estimation of logistic regression models: theory and implementation*“ [9].

Olgu $\mathbf{y}_i^T = (y_{i1}, \ldots, y_{iq}) \sim M(n_i, \pi_i), \quad i = 1, \ldots, n$ multinoomjaotusest, mille võimalike tasemete arv on $k = q + 1$ ja tõenäosusfunktsioon

$$P_{n_i}(\mathbf{y_i}) = \frac{n_i!}{y_{i1}! \ \ldots \ y_{iq}! \, (n_i - y_{i1} - \ldots - y_{iq})!} \pi_{i1}^{y_{i1}} \ \ldots \ \pi_{iq}^{y_{iq}} \cdot \\ \cdot (1 - \pi_{i1} - \ldots - \pi_{iq})^{n_i - y_{i1} - \ldots - y_{iq}}.$$

Kuna jagatises $\dfrac{n_i!}{y_{i1}! \ \ldots \ y_{iq}! \, (n_i - y_{i1} - \ldots - y_{iq})!}$ ei ole hinnatavaid tõenäosusi π_{ir}, siis võib seda vaadelda konstandina ja suurima tõepära funktsioon avaldub kujul

$$L(\boldsymbol{\beta}) \simeq \prod_{i=1}^{n} \pi_{i1}^{y_{i1}} \ \ldots \ \pi_{iq}^{y_{iq}} (1 - \pi_{i1} - \ldots - \pi_{iq})^{n_i - y_{i1} - \ldots - y_{iq}} = \\ = \prod_{i=1}^{n} \prod_{r=1}^{q} \left(\frac{\pi_{ir}}{\pi_{ik}} \right)^{y_{ir}} \cdot \pi_{ik}^{n_i}.$$

Asendades π_{ir} ja π_{ik} vastavalt valemite 2 ja 3 järgi, saame

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{r=1}^{q} e^{y_{ir} \sum_{j=0}^{m} \beta_{rj} x_{ij}} \cdot \left(1 + \sum_{r=1}^{q} e^{\sum_{j=0}^{m} \beta_{rj} x_{ij}} \right). \tag{4}$$

Suurima tõepära hinnangute leidmiseks on vaja eelnev funktsioon maksimeerida. Valemist 4 naturaallogaritmi võtmisel saame log-tõepära funktsiooni Kuna logaritm on monotoonne funktsioon, siis piisab log-tõepära funktsiooni maksimumkohtade leidmisest.

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{r=1}^{q} \left(y_{ir} \sum_{j=0}^{m} \beta_{rj} x_{ij} \right) - n_i \ln \left(1 + \sum_{r=1}^{q} e^{\sum_{j=0}^{m} \beta_{rj} x_{ij}} \right).$$

Funktsiooni maksimeerimiseks $\boldsymbol{\beta}$ suhtes lahendame nn tõepäravõrrandid $\dfrac{\partial l(\boldsymbol{\beta})}{\partial \beta_{rj}} = 0$. Tegemist on mittelineaarse võrrandisüsteemiga, mida lahendatakse iteratiivselt kasutatades selleks tavaliselt Newtoni-Raphsoni meetodit.

2.2 Peakomponentide analüüs

Käesolev peatükk põhineb Tartu Ülikooli matemaatika ja statistika instituudi dotsendi Imbi Traadi mitmemõõtmelise analüüsi loengukonspektil [10], kui pole viidatud teisiti.

Olgu meil p tunnust ning neile vastav juhuslik vektor $\mathbf{X}^T = (X_1, X_2, \ldots, X_p)$ keskväärtusega $E(\mathbf{X})$ ja dispersioonimaatriksiga

$$D(\mathbf{X}) = \boldsymbol{\Sigma} = E\left((\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})^T\right) \ : p \times p.$$

Peakomponentide analüüsi eesmärk on leida omavahel mittekorreleeritud uued tunnused (peakomponendid) $P_1, \ldots, P_p$, mis on esialgsete tunnuste $X_1, \ldots, X_p$ lineaarkombinatsioonid

$$P_i = \mathbf{a_i}^T \mathbf{X} = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p, \quad i = 1, \ldots, p,$$

kordajate vektoriga $\mathbf{a_i}^T = (a_{i1}, \ldots, a_{ip})$ nii, et komponendil P_1 on maksimaalne võimalik dispersioon, komponendil P_2 suuruselt järgmine dispersioon jne.

Seega esmalt otsime vektorit $\mathbf{a_1}$ selliselt, et dispersioon

$$D(P_1) = D(\mathbf{a_1}\mathbf{X}) = \mathbf{a_1}^T \boldsymbol{\Sigma} \mathbf{a_1}$$

oleks maksimaalne. On selge, et eelnevat suurust on võimalik lõpmatult suurendada, kor-

rutades seda läbi mingi konstandiga. Et määrata otsitav vektor üheselt, seatakse talle normeerituse kitsendus ehk $\mathbf{a_1}^T\mathbf{a_1} = 1$.

Olgu dispersioonimaatriksi $\mathbf{\Sigma}$ omaväärtused $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ ning neile vastavad omavektorid $\gamma_\mathbf{1}, \gamma_\mathbf{2}, \ldots, \gamma_\mathbf{p}$. Selgub, et selliseks vektoriks $\mathbf{a_1}$ sobib võtta dispersioonimaatriksi $\mathbf{\Sigma}$ omaväärtusele λ_1 vastav omavektor $\gamma_\mathbf{1}$ ehk $P_1 = \gamma_\mathbf{1}{}^T\mathbf{X}$ ning siis on peakomponendi P_1 dispersioon $DP_1 = \lambda_1$, mis on suurim võimalik. (Tõestus Johnsoni ja Wicherni õpikus [11] ja I. Traadi konspektis [10]). Teise peakomponendi defineerimiseks sobib seega kasutada omaväärtusele λ_2 vastavat omavektorit ning $P_2 = \gamma_\mathbf{1}\mathbf{X}$ ja $DP_2 = \lambda_2$. Kuna dispersioonimaatriksi omavektorid on ortogonaalsed, siis on peakomponendid P_1 ja P_2 mittekorreleeritud.

Lähtetunnuste kogudispersioon $DX_1 + \ldots + DX_p$ on maatriksi $\mathbf{\Sigma}$ jälg. Ühtlasi on maatriksi jälg ka tema omaväärtuste summa. Seega

$$DX_1 + \ldots + DX_p = tr(\mathbf{\Sigma}) = \lambda_1 + \ldots + \lambda_p,$$

mis tähendab, et lähtetunnuste kogudispersioon on võrdne peakomponentide kogudispersiooniga, kusjuures iga järgmise peakomponendi dispersioon on maksimaalne võimalik. Kokkuvõttes valem i-nda peakomponendi määramiseks on

$$P_i = \gamma_\mathbf{i}{}^T\mathbf{X}$$

ning tema dispersioon on

$$DP_i = \gamma_\mathbf{i}{}^T\mathbf{\Sigma}\gamma_\mathbf{i} = \lambda_i.$$

Peakomponendi tähtsuse näitajaks on osakaal

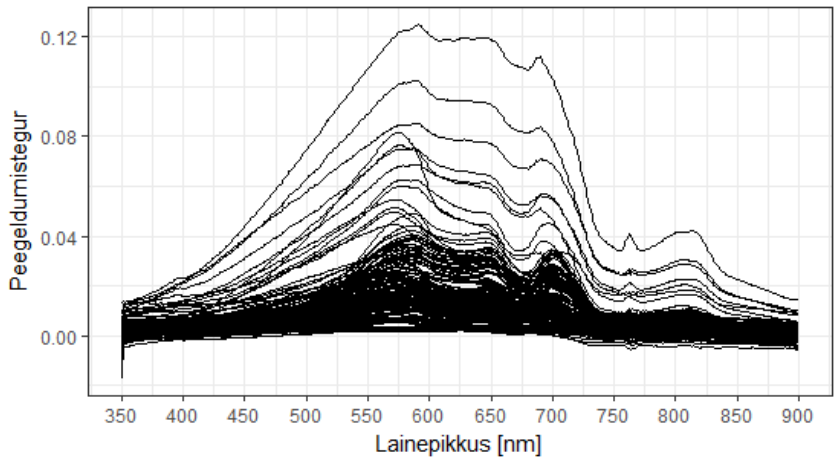
$$\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i},$$

mis näitab, kui suure osa tunnuste koguvarieeruvusest kirjeldab i-s peakomponent.

3 Veekogude klassifitseerimise mudel

3.1 Andmestiku ülevaade

Veekogude peegeldumisspektreid sisaldavas andmestikus on 180 *in situ* mõõtmistulemuse andmed (visualiseeritud joonisel 1), mis on kogutud aastatel 2006 ning 2012–2016 Tartu Observatooriumi veekogude kaugseire töörühma poolt kolmest spektromeetrist koosneva süsteemiga TriOS Ramses. Nendest mõõtmistest 83 pärinevad Peipsi järvest, 31 Põhjamere kaguosas asuvast Waddenzeest, 30 Võrtsjärvest, 3 Rootsi ranniku lähedal asuvast Läänemere osast ning ülejäänud mõõtmised pärinevad Eesti väikejärvedest. Uuritav tunnus on veetüüp, millel on vastavalt Uudeberg jt [7] klassifikatsioonile viis võimalikku väärtust: *Selge*, *Mõõdukas*, *Sogane*, *Väga sogane* ning *Pruun* (alajaotus 1.3).



Joonis 1. Kõigi andmestikus olevate *in situ* vaatluste peegeldumisspektrid

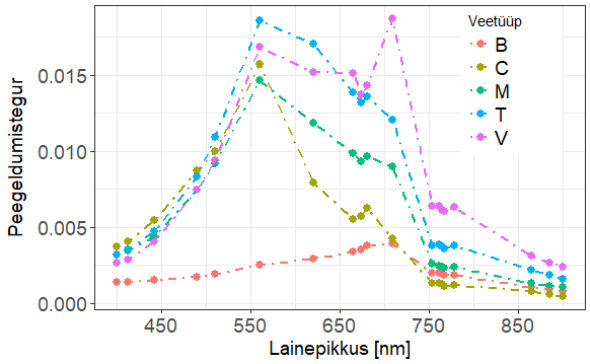
Kuigi satelliitsensoril OLCI on 21 kanalit, siis vaid 19 neist asuvad Ramsese kalibratsiooniga kattuvatel lainepikkustel. Nendeks lainepikkusteks on 400 nm, 412,5 nm, 442,5 nm, 490 nm, 510 nm, 560 nm, 620 nm, 665 nm, 673,75 nm, 681,25 nm, 708,75 nm, 753,75 nm, 761,25 nm, 764,375 nm, 767,5 nm, 778,75 nm, 865 nm, 885 nm, 900 nm. Seetõttu on Tartu Observatooriumi veekogude kaugseire töörühma poolt välitöödel kogutud hüperspektraalsete andmete põhjal leitud peegeldumistegurid ümber arvutatud OLCI kanalitega kattuvatele lainepikkustele. Analüüsitavateks tunnusteks on seega 19 erineval lainepikkusel leitud peegeldumistegurit. Peegeldumisteguri arvutamise metoodika tõttu

võivad mõned peegeldumisteguri väärtused olla ka negatiivsed (alajaotus 1.2).

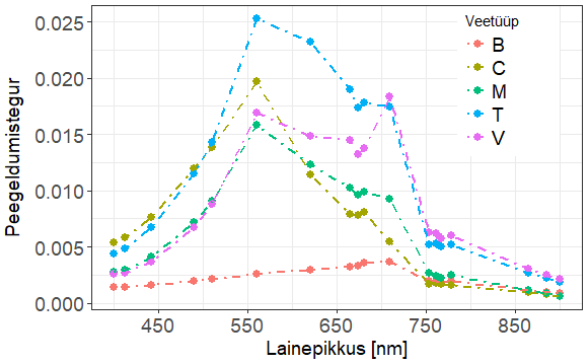
3.2 Kirjeldav analüüs

Bakalaureusetöö aluseks olnud andmestiku uurimisel selgus, et vaadeldud veekogude peegeldumisteguri maksimaalne väärtus on 0,11888 lainepikkusel 620 nm ning minimaalne väärtus on -0,00529 lainepikkusel 900 nm. Peegeldumisteguri keskmine väärtus on 0,00813 ning mediaan on 0,00473. Seega enam kui pooled uuritavatest spektritest on väga madalate väärtustega, mis on ka Eesti tumedate veekogude puhul ootuspärane tulemus. Veekogudel mõõdetud peegeldumisspektritest 159 saavutavad maksimumi lainepikkustel 560 või 620 nanomeetrit ning ülejäänud 21 spektrit saavutavad maksimumi lainepikkustel 681,25 või 708,75 nanomeetrit.

Kõigist vaatlustest kuuluvad 34 *Selgesse*, 29 *Mõõdukasse*, 94 *Sogasesse*, 14 *Väga sogasesse* ning 9 *Pruuni* veetüüpi. *Selgesse* veetüüpi kuuluvad peamiselt Waddenzee ja Läänemere mõõtepunktid ning lisaks ka üksikud Eesti väikejärved nagu Saadjärv ja Koorküla Valgjärv. *Mõõduka* ja *Väga sogase* veetüübi moodustavad suures osas Peipsi järve erinevad mõõtepunktid. *Pruuni* veetüübi mõõtmised on pärit Eesti tumedatest väikejärvedest (nt Leego järv ja Valguta Mustjärv). *Sogasesse* veetüüpi kuuluvad mõõtmised pärinevad suures osas Võrtsjärvest ja Peipsi järvest. Samuti on *Sogase* veetüübiga ka mõningad Waddenzee rannikuäärsed kohad ja mitmed Eesti väikejärved, nagu näiteks Pühajärv ja Verevi järv.



Joonis 2. Veetüüpide spektrite mediaanid



Joonis 3. Veetüüpide spektrite keskmised

Joonisel 2 on toodud veetüüpide mediaanspektrid. On näha, et *Selgele* veetüübile (C) on

omane peegeldumisteguri väärtuste kiire langus lainepikkuse suurenemisel pärast maksimumi saavutamist lainepikkusel 560 nanomeetrit. *Mõõduka* (M) ja *Sogase* (T) veetüübi mediaanspektrid saavutavad samuti maksimumi lainepikkusel 560 nanomeetrit, kuid sellele järgnev peegeldumisteguri langemine ei ole nii järsk kui *Selge* veetüübi puhul. Samuti on *Sogase* veetüübi mediaanspekter veidi kõrgemate väärtustega kui *Mõõduka* tüübi spekter. *Väga sogase* veetüübi (V) peegeldumisteguri mediaanväärtuste maksimum on lainepikkusel 708,75 nanomeetrit. *Pruuni* veetüübi (B) puhul saavutab peegeldumisteguri maksimumi samuti pikematel lainepikkustel 708 nanomeetri ümber, kuid võrreldes teiste veetüüpidega on peegeldumisteguri väärtused kogu spektri ulatuses tunduvalt madalamad. Sarnane tendents iseloomustab ka veetüüpide keskmiseid peegeldumisspektreid (joonis 3).

3.3 Statistiline mudel

Töö peamiseks eesmärgiks oli, tuginedes maapealsete mõõtmiste andmetele, luua mudel, mis võimaldaks satelliidiandmete põhjal veekogusid klassifitseerida. Uuritaval tunnusel veetüüp on viis võimalikku väärtust ning seega on tegemist viie tasemega tunnusega, mis on multinoomjaotusega. Multinomiaalse logistilise regressiooniga saame veekogu peegeldumisspektri alusel hinnata igale tasemele kuulumise tõenäosust ehk määrata, millise tõenäosusega kuulub veekogu tüüpidesse *Selge*, *Mõõdukas*, *Sogane*, *Väga sogane* ja *Pruun*.

Analüüsitavate tunnuste tugevale seotusele viitab juba andmete olemus. Peegeldumisteguri väärtus mingil lainepikkusel on tugevalt seotud peegeldumisteguri väärtustega eelnevatel ja järgnevatel lainepikkustel. Korrelatsioonimaatriksi uurimisel selgub, et vähim korrelatsioon on ligikaudu 0,69 ning see esineb lainepikkuste 490 ning 900 vahel. Ligikaudu poolte lainepikkuste paaride korrelatsioonid on suuremad kui 0,9 ning suurimad korrelatsioonid küündivad üle 0,99 (lisa 2). Niivõrd suurte korrelatsioonide puhul kirjeldavad analüüsitavad tunnused alati rohkem üksteist kui uuritavat tunnust. Seetõttu on käesolevas töös kasutusele võetud peakomponentide analüüs.

3.3.1 Peakomponentide analüüsi rakendamine

Peakomponentide analüüsi rakendamisel on võimalik saada omavahel mittekorreleeritud uued tunnused, mida multinomiaalse logistilise regressioonimudeli loomisel lainepikkuste (argumentide) asemel kasutada. Enne peakomponentide loomist on peegeldumistegurid erinevatel lainepikkustel standardiseeritud nii, et iga lainepikkuse jaoks oleks peegeldumistegurite keskmine 0 ja standardhälve 1.

Peakomponentide analüüsi tulemusena selgus, et esimene peakomponent PC1 kirjeldab koguvarieeruvusest 88,3% ning teine peakomponent PC2 kirjeldab 6,6%. Seega kaks esimest peakomponenti kirjeldavad kokku juba peaaegu 95% valimi andmete varieeruvusest. Peakomponendid PC3 ja PC4 kirjeldavad vastavalt 3,5% ja 0,9% koguvarieeruvusest, mis tähendab, et esimese nelja peakomponendi abil on võimalik kirjeldada üle 99% kogu andmetes esinevast varieeruvusest. Sellest lähtudes on edaspidises analüüsis kasutatud vaid nelja esimest peakomponenti (lisa 3).

Lisas 4 on toodud kaalud leitud peakomponentide jaoks. Esimese peakomponendi PC1 kaalud iga esialgse lainepikkuse jaoks on omavahel suhteliselt sarnased, varieerudes vahemikus 0,21–0,24. Üldiselt, mida suuremad on peegeldumisspektri väärtused, seda suurem on ka peakomponendi PC1 väärtus.

Teise peakomponendi PC2 kaalud on lainepikkustel 400–681,25 nm negatiivsed ning lainepikkustel 708,75–900 nm positiivsed, mis tähendab, et kõrgemad peegeldumisteguri väärtused lühematel lainepikkustel muudavad peakomponendi PC2 väärtust väiksemaks ning kõrgemad peegeldumisteguri väärtused pikematel lainepikkustel muudavad peakomponendi PC2 väärtust suuremaks. Samuti võib täheldada, et keskmistel lainepikkustel on PC2 kaalude väärtused nullile lähemal ($<|0{,}05|$) kui äärmistel lainepikkustel($>|0{,}3|$). Seega mõjutab peegeldumisspektri keskosa peakomponendi PC2 väärtust vähem kui äärmised osad.

Kolmanda peakomponendi PC3 jaoks paiknevad mõjukamad kaalud spektri äärtes ja keskel. Lühematel lainepikkustel kui 450 nanomeetrit ja pikematel lainepikkustel kui 865 nanomeetrit on kaalud umbes –0,3 ning lainepikkustel 620–708,75 nm on kaalud umbes +0,3, mis tähendab, et suuremad peegeldumisteguri väärtused spektri keskosas muuda-

vad peakomponendi PC3 väärtust suuremaks ning suuremate peegeldumisteguri väärtuste paiknemine spektri äärtes muudavad peakomponendi PC3 väärtust väiksemaks.

Neljanda peakomponendi PC4 korral on kaalud lainepikkuste 865–900 nm ning 490–620 nm jaoks negatiivsed ehk suuremad peegeldumisteguri väärtused neil lainepikkustel muudavad peakomponendi PC4 väärtust väiksemaks ning suuremad peegeldumisteguri väärtused ülejäänud lainepikkustel muudavad peakomponendi PC4 väärtust suuremaks.

3.3.2 Multinomiaalne logistiline regressioonimudel

Multinomiaalse logistilise regressioonimudeli loomisel on aluseks andmestik, kus on 180 vaatlust. Uuritav tunnus on veetüüp viie võimaliku väärtusega *Selge*, *Mõõdukas*, *Sogane*, *Väga sogane* ja *Pruun* ning kirjeldavateks tunnusteks on eelmises alapeatükis leitud peakomponendid. Kuna iga järgmine peakomponent kirjeldab maksimaalse võimaliku osa allesolevast varieeruvusest, siis on mudeli koostamisel alustatud ainult ühe peakomponendiga mudelist ning seejärel hakatud peakomponente ükshaaval järjest juurde lisama.

Mudeli baastasemeks on valitud *Sogane* veetüüp, sest sellesse tüüpi kuulub veidi enam kui 50% andmestikus olevatest vaatlustest. Määrates iga vaatluse jaoks prognoositud veetüübiks maksimaalse prognoositud tõenäosusega tüübi, on võimalik hinnata, kui suure osa vaatlustest mudel õigesti prognoosib. Toetudes teadmisele, et neli esimest peakomponenti kirjeldavad enam kui 99% kogu andmetes esinevast varieeruvusest ning et nelja peakomponendiga mudel prognoosib õigesti 81% vaatlustest ja rohkemate peakomponentide lisamine mudelit palju ei paranda, on lõplikuks mudeliks valitud järgmine mudel (lisa 5):

$$\ln \frac{P(Pruun)}{P(Sogane)} = -167,013 - 39,475 \cdot PC1 + 77,547 \cdot PC2 - 20,813 \cdot PC3 + 97,816 \cdot PC4$$

$$\ln \frac{P(Selge)}{P(Sogane)} = -8,217 - 2,145 \cdot PC1 - 4,056 \cdot PC2 - 6,992 \cdot PC3 - 4,459 \cdot PC4$$

$$\ln \frac{P(Mõõdukas)}{P(Sogane)} = -1,230 - 0,329 \cdot PC1 - 0,396 \cdot PC2 - 0,849 \cdot PC3 - 0,388 \cdot PC4$$

$$\ln \frac{P(Väga\ sogane)}{P(Sogane)} = -29,921 - 4,581 \cdot PC1 + 22,284 \cdot PC2 + 5,059 \cdot PC3 + 37,218 \cdot PC4$$

Kõigi veetüüpide jaoks on nii vabaliige kui ka esimese peakomponendi PC1 kordaja negatiivsed. Vabaliikme negatiivsus tähendab, et üldiselt on veekogul suurem šanss kuuluda *Sogasesse* veetüüpi kui ükskõik millisesse muusse ülejäänud neljast veetüübist. Peakomponendi PC1 negatiivsed kordajad näitavad, et esimese peakomponendi väärtuste suurenemisel ehk peegeldumistegurite väärtuste suurenemisel suureneb ka šanss kuuluda *Sogasesse* veetüüpi. Peakomponendi PC2 kordajad on *Pruuni* ja *Väga sogase* veetüübi jaoks positiivsed, mis tähendab, et peakomponendi PC2 suurenemisel suureneb šanss kuuluda vastavalt *Pruuni* või *Väga sogasesse* veetüüpi. Seega, mida madalamad on peegeldumisspektri väärtused lühematel lainepikkustel ja mida kõrgemad on peegeldumisspektri väärtused pikematel lainepikkustel, seda suurem on šanss kuuluda *Pruuni* või *Sogasesse* veetüüpi. Samuti, kuna *Selge* ja *Mõõduka* veetüübi jaoks on peakomponendi PC2 kordajad negatiivsed, siis seda väiksem on šanss kuuluda *Selgesse* või *Mõõdukasse* veetüüpi. Peakomponentide PC3 ja PC4 kordajad on *Selge* ja *Mõõduka* veetüübi jaoks negatiivsed, seega nende väiksem väärtus suurendab šanssi kuuluda *Selgesse* või *Mõõdukasse* veetüüpi. *Väga sogase* veetüübi jaoks on peakomponentide PC3 ja PC4 kordajad positiivsed ja seega nende suurem väärtus suurendab šanssi kuuluda *Väga sogasesse* veetüüpi. Peakomponendi PC3 kordaja on *Pruuni* veetüübi jaoks negatiivne ja peakomponendi PC4 kordaja positiivne. Seega peakomponendi PC3 väiksem väärtus ning peakomponendi PC4 suurem väärtus suurendavad šanssi kuuluda *Pruuni* veetüüpi.

Valides olulisuse nivooks $\alpha = 0,1$, võib tähele panna, et *Pruuni* veetüübi jaoks ei ole ükski mudeli kordaja statistiliselt oluline. See tähendab, et mudelis ei eristu *Pruun* veetüüp *Sogasest* ning on ilmselt tingitud asjaolust, et *Pruuni* veetüübi vaatluseid on andmestikus väga vähe. *Selge* veetüübi jaoks on mudeli kõik kordajad statistiliselt olulised. *Mõõduka* veetüübi jaoks ei ole peakomponentide PC2 ja PC4 kordajad olulised ning *Väga sogase* veetüübi jaoks on olulised kõik kordajad peale peakomponendi PC3 kordaja.

Mudeli kasutamise illustreerimiseks vaatame Peipsi järve andmeid. Algselt olid Peipsi järve 83 mõõtepunkti hulgas esindatud kõik viis veetüüpi. *Pruuni* veetüüpi kuulus 1,2% vaatlustest, *Selgesse* veetüüpi 2,4%, *Mõõdukasse* veetüüpi 31,3%, *Sogasesse* veetüüpi 48,2% ning *Väga sogasesse* tüüpi kuulus 16,9% vaatlustest. Mudeli rakendamisel prognoositud klassid jaotusid aga veidi teisiti. *Pruuni* ja *Väga sogase* veetüübi puhul jäid osakaalud sa-

maks, olles vastavalt 1,2% ja 16,9%. *Selgesse* veetüüpi prognoositi 4,8% vaatlustest. Mudel aga ei prognoosinud ühegi vaatluse tüübiks *Mõõdukat* veetüüpi ning suuremas osas klassifitseeris need hoopis *Sogase* veetüübi alla. Seetõttu on prognoositud veetüüpidest 77,1% klassifitseeritud *Sogasesse* veetüüpi, mida on oluliselt rohkem kui algses jaotumises.

Võrdluseks võib vaadata Võrtsjärve, kus olid algselt kõik vaatlused *Sogasest* veetüübist ning mudel kõik sinna ka klassifitseeris.

3.4 Veekogude klassifitseerimine mudeli põhjal

Määrates koostatud mudeli poolt prognoositud tõenäosuste järgi vaatlusele suurima prognoositud tõenäosusega veetüübi, saame hinnata, kui suure osa vaatlustest mudel õigesti prognoosib.

Esimene katsetatud mudel, milles oli lisaks vabaliikmele vaid esimene peakomponent PC1, jagas veekogud kolme tüübi vahel: *Pruun*, *Mõõdukas* ja *Sogane*. Mudel klassifitseeris õigesti 51% vaatlustest. Õigesti klassifitseeriti vaid *Sogase* veetüübi mõõtmised, 94-st *Sogase* tüübi vaatlusest 92 klassifitseeriti *Sogase* tüübi alla (tabel 1).

Tabel 1. Ühe peakomponendiga mudeli prognooside võrdlus tegelikega

tegelik / mudel	*Pruun*	*Selge*	*Mõõdukas*	*Sogane*	*Väga sogane*
Pruun	0	1	0	2	1
Mõõdukas	0	1	0	0	0
Sogane	9	32	29	92	13

Teine mudel, kus oli lisaks vabaliikmele ja esimesele peakomponendile PC1 ka teine peakomponent PC2, jagas veekogud nelja tüübi vahel: *Pruun*, *Selge*, *Sogane* ja *Väga sogane*. Mudel klassifitseeris õigesti 8 *Pruuni*, 16 *Selge*, 86 *Sogase* ja 7 *Väga sogase* veetüübi vaatlust. Suurema osa *Selge* ja *Mõõduka* veetüübi vaatlustest klassifitseeris kahe peakomponendiga mudel *Sogase* veetüübi alla. Kokku klassifitseeris selline mudel õigesti 65% vaatlustest (tabel 2).

Mudel, kuhu oli lisatud ka kolmas peakomponent PC3, eristas samuti *Pruuni*, *Selge*, *Sogase* ja *Väga sogase* veetüübi ning klassifitseeris õigesti 76% vaatlustest. Enamik *Pruuni*, *Selge*

Tabel 2. Kahe peakomponendiga mudeli prognooside võrdlus tegelikega

mudel \ tegelik	*Pruun*	*Selge*	*Mõõdukas*	*Sogane*	*Väga sogane*
Pruun	8	0	1	1	0
Selge	0	16	4	6	1
Sogane	1	18	23	86	6
Väga sogane	0	0	1	1	7

ja *Sogase* veetüübi vaatlustest olid klassifitseeritud õigesti, kuid suurem osa *Mõõduka* ja *Väga sogase* veetüübi vaatlustest klassifitseeriti *Sogase* veetüübi alla (tabel 3).

Tabel 3. Kolme peakomponendiga mudeli prognooside võrdlus tegelikega

mudel \ tegelik	*Pruun*	*Selge*	*Mõõdukas*	*Sogane*	*Väga sogane*
Pruun	9	0	0	1	0
Selge	0	32	4	2	0
Sogane	0	2	23	90	8
Väga sogane	0	0	2	1	6

Nelja peakomponendiga lõplik mudel jagas veekogud samuti vaid nelja eelnevalt mainitud klassi vahel, kuid kokku klassifitseeris õigesti 81% vaatlustest. *Mõõduka* veetüübi 29 vaatlusest klassifitseeriti 24 *Sogase*, 4 *Selge* ja 1 *Väga sogase* veetüübi alla. *Selge*, *Sogase*, *Pruuni* ja *Väga sogase* veetüübi vaatlused klassifitseeris mudel enamjaolt õigesti (tabel 4).

Tabel 4. Nelja peakomponendiga mudeli prognooside võrdlus tegelikega

mudel \ tegelik	*Pruun*	*Selge*	*Mõõdukas*	*Sogane*	*Väga sogane*
Pruun	9	0	0	0	0
Selge	0	32	4	2	0
Sogane	0	2	24	92	1
Väga sogane	0	0	1	0	13

Lisades mudelisse ka viienda peakomponendi PC5, jagas mudel veekogud küll viide klassi, kuid *Mõõdukasse* veetüüpi klassifitseeriti vaid kaks *Sogase* veetüübiga vaatlust. Samuti langes õigesti prognoositud vaatluste arv 79% peale. Kuuenda peakomponendi lisamisega tõusis õigesti klassifitseeritavate vaatluste osa 82 protsendini ja prognoositi viis veetüüpi, kuid 25 *Mõõduka* veetüübi vaatlust 29-st läks ikkagi *Sogase* veetüübi alla. Lisades mu-

delisse ka seitsmenda peakomponendi PC7, klassifitseeris mudel õigesti 84% vaatlustest. *Mõõduka* veetüübi vaatlustest klassifitseerusid enamikud ikka *Sogase* veetüübi alla (lisa 6).

Toodud prognooside analüüs kinnitab esialgset mudeli valikut, kus on kasutusel neli esimest peakomponenti. Antud nelja komponendiga mudel ei erista *Mõõduka* veetüübi vaatluseid (klassifitseerides need peamiselt *Sogase* veetüübi alla), kuid ka rohkemate peakomponentide arvuga mudelid määravad suurema osa *Mõõduka* veetüübi vaatlustest ikkagi *Sogase* veetüübi alla ning suudavad õigesti klassifitseerida vaid üksikud *Mõõduka* veetüübi vaatlused.

Kokkuvõte

Bakalaureusetöö eesmärk oli luua mudel Eestile ja selle lähiümbruskonnale omaste optiliselt keeruliste veekogude klassifitseerimiseks satelliidiandmetelt. Kasutatavad andmed on kogutud Tartu Ülikooli Tartu Observatooriumi veekogude kaugseire töörühma teadlaste poolt aastatel 2006 ning 2012–2016. Andmestikus esineva tugeva tunnustevahelise korrelatsiooni eemaldamiseks on enne mudeli loomist rakendatud peakomponentide analüüsi. Loodud peakomponentide põhjal on koostatud multinomiaalne logistiline regressioonimudel veekogude klassifitseerimiseks.

Töö aluseks võetud Uudeberg jt [7] klassifikatsioon jagab Eesti ja selle lähiümbruskonna järved ja rannikuveed viide klassi: *Pruun*, *Selge*, *Mõõdukas*, *Sogane*, *Väga sogane*. Töö käigus valminud viie tasemega multinomiaalne logistiline regressioonimudel hindab veekogu kohal mõõdetud peegeldumisspektri jaoks, millise tõenäosusega mõõtetulemus igasse viiest veetüübist kuulub. Maksimaalse hinnatud tõenäosusega veetüüp määratakse prognoositud veetüübiks.

Saadud mudelis on argumentideks peakomponentide analüüsi osas hinnatud neli esimest peakomponenti ning mudel hindab õigesti 81% andmestikes olevatest vaatlustest. *Selge*, *Sogase*, *Väga sogase* ja *Pruuni* veetüübi vaatlused klassifitseeritakse suures osas õigesti, kuid *Mõõdukasse* klassi ei prognoosi mudel ühtegi vaatlust. Suur osa *Mõõduka* veetüübi vaatlustest klassifitseeritakse selle asemel *Sogasesse* veetüüpi st mudel ei erista *Mõõdukat* veetüüpi *Sogasest*.

Käesoleva bakalaureusetöö tulemusena saadud mudel on kasutatav veekogude kaugseirega seotud teadustöö arendamisel, kuid vajab täiendavat analüüsi *Mõõduka* veetüübi eristamiseks. Veekogude klassifitseerimine võimaldab efektiivsemalt kasutusele võtta rahvusvaheliselt tunnustatud optiliselt aktiivsete ainete sisalduse hindamiseks loodud algoritme.

Viited

[1] Arst, H. (2003). *Optical Properties and Remote Sensing of Multicomponental Water Bodies.* Berlin: Springer-Verlag Berlin Heidelberg.

[2] IOCCG (2000). *Remote Sensing of Ocean Colour in Coastal, and Other Optically-Complex, Waters.* Sathyendranath, S. (toim.), Reports of the International Ocean-Colour Coordinating Group, No. 3, IOCCG. Dartmouth, Canada.

[3] Copernicus: Sentinel-3. https://directory.eoportal.org/web/eoportal/satellite-missions/c-missions/copernicus-sentinel-3 (17.04.2018).

[4] Gordon, H., Brown, O. B., Jacobs, M. M. (1975). Computed Relationships Between the Inherent and Apparent Optical Properties of a Flat Homogeneous Ocean. *Applied optics. 14.* 417-27. 10.1364/AO.14.000417.

[5] Morel, A., Prieur, L. (1977). Analysis of variations in ocean color, *Limnology and Oceanography*, 4, doi: 10.4319/lo.1977.22.4.0709

[6] Reinart, A., Herlevi, A., Arst, H., Sipelgas, L. (2003). Preliminary optical classification of lakes and coastal waters in Estonia and south Finland. *Journal of Sea Research, 49,* 357-366. doi:10.1016/S1385-1101(03)00019-4

[7] Uudeberg, K., Põru, G., Ansko, I., Ansper, A., Ligi, M. (2017). Estimation of the lakes optical water types from satellites' images. Poster HIGHROC Science Conference, Brüssel.

[8] Tutz, G. (2011). *Regression for Categorical Data* (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511842061

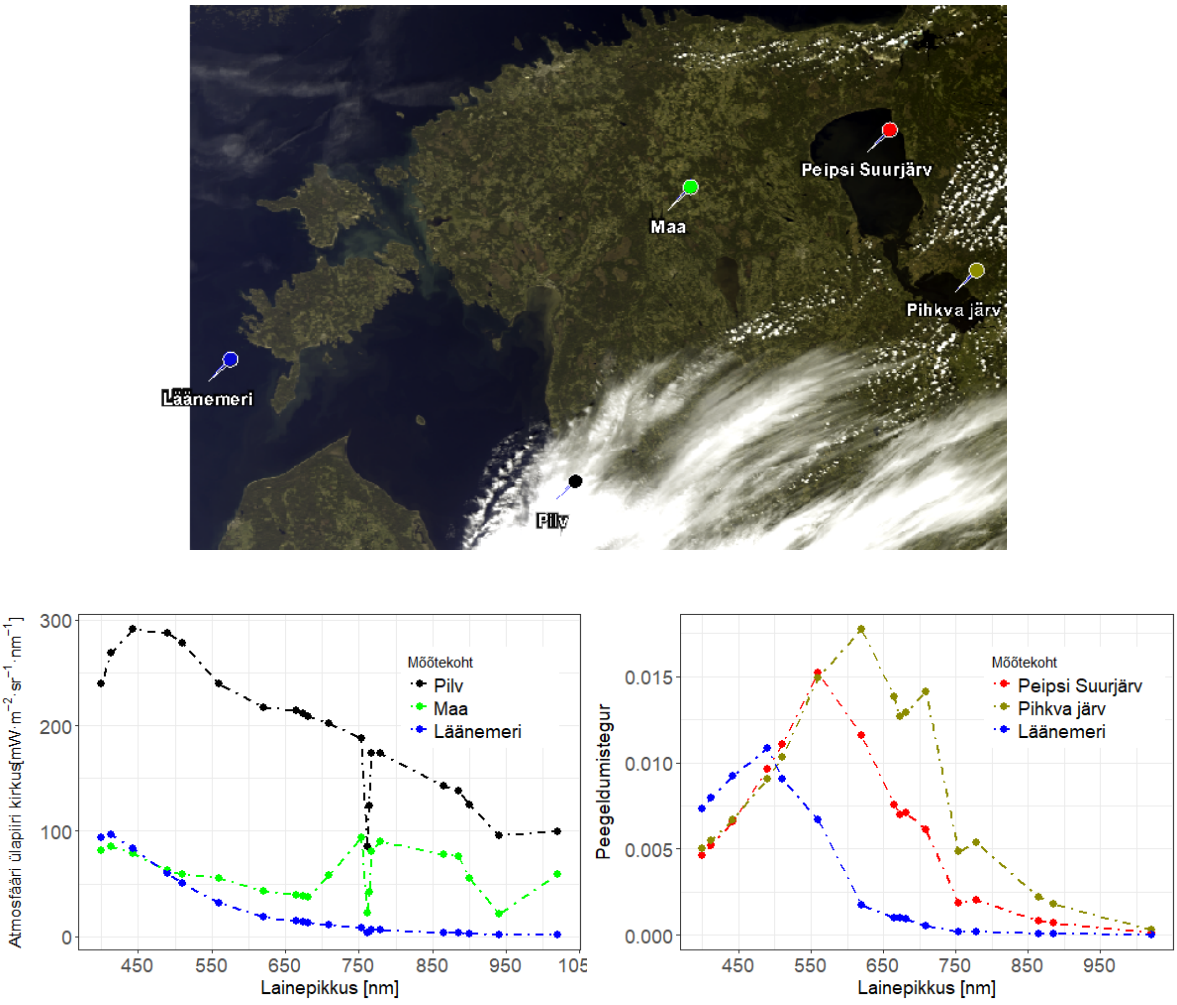
[9] Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: theory and implementation. https://czep.net/stat/mlelr.pdf (11.04.2018).

[10] Traat, I. (2016). *Mitmemõõtmeline analüüs.* Loengukonspekt.

[11] Johnson, R. A. (2007). *Applied Multivariate Statistical Analysis.* (Sixth Edition) New Jersey: Wiley.

Lisad

Lisa 1. Satelliidipildilt saadav info



Joonis 4. ESA satelliidi Sentinel-3 sensori OLCI Level-1 töötlusega RGB-pilt Eestist 14.06.2016 ning pildilt saadav spektraalne info

Joonise 4 alumine vasakpoolne graafik kujutab satelliidipildil märgitud pilve, maa ja Läänemere pikslite atmosfääri ülapiiri kirkuse spektrite kujusid lainepikkustel 400-1020 nm. Spektraalne info on saadud sensori OLCI Level-1 töötlusega pildilt, abstsissteljel paiknevad OLCI kanalitele vastavad lainepikkused ja ordinaatteljel on märgitud atmosfääri ülapiiri kirkus (*TOA (Top of the Atmosphere) radiance*). Jooniselt on näha, et veekogu kohalt saadav signaal on oluliselt nõrgem kui pilve ja maa kohalt mõõdetava kiirguse oma.

Joonise 4 alumisel parempoolsel graafikul on toodud Suurjärve (Peipsi järve põhjaosa), Pihkva järve ja Läänemere pikslite vastavate peegeldumisspektrite kujud. Spektraalne info on saadud sensori OLCI C2RCC töötlusega pildilt ning abstsissteljel paiknevad OLCI kanalitele vastavad lainepikkused ning ordinaatteljel on märgitud peegeldumistegur. Graafikult näeme, et erinevat tüüpi veed on erineva peegeldumisspektri kuju ja väärtustega.

Lisa 2. Kirjeldavate tunnuste korrelatsioonimaatriks

Tabel 5. Korrelatsioonimaatriks kirjeldavate tunnuste kohta

	400	412,5	442,5	490	510	560	620	665	673,75	681,25	708,75	753,75	761,25	764,375	767,5	778,75	865	885	900
400	1	0,9966	0,9628	0,9070	0,8891	0,8320	0,7922	0,7729	0,7872	0,7883	0,7074	0,7606	0,7733	0,7795	0,7794	0,7671	0,7627	0,7423	0,7088
412,5	0,9966	1	0,9809	0,9364	0,9208	0,8642	0,8244	0,8058	0,8209	0,8218	0,7332	0,7809	0,7952	0,8012	0,8005	0,7882	0,7790	0,7561	0,7190
442,5	0,9628	0,9809	1	0,9849	0,9746	0,9170	0,8733	0,8577	0,8756	0,8756	0,7682	0,8032	0,8183	0,8245	0,8243	0,8127	0,7892	0,7596	0,7110
490	0,9070	0,9364	0,9849	1	0,9978	0,9529	0,8995	0,8813	0,8985	0,8984	0,7853	0,8012	0,8167	0,8226	0,8226	0,8120	0,7804	0,7480	0,6923
510	0,8891	0,9208	0,9746	0,9978	1	0,9691	0,9175	0,8974	0,9117	0,9116	0,8061	0,8100	0,8244	0,8300	0,8304	0,8208	0,7862	0,7534	0,6970
560	0,8320	0,8642	0,9170	0,9529	0,9691	1	0,9626	0,9307	0,9301	0,9307	0,8715	0,8285	0,8374	0,8419	0,8439	0,8388	0,7957	0,7645	0,7111
620	0,7922	0,8244	0,8733	0,8995	0,9175	0,9626	1	0,9895	0,9839	0,9844	0,9539	0,8982	0,8992	0,9008	0,9042	0,9052	0,8357	0,7963	0,7401
665	0,7729	0,8058	0,8577	0,8813	0,8974	0,9307	0,9895	1	0,9977	0,9973	0,9686	0,9299	0,9282	0,9286	0,9323	0,9349	0,8615	0,8200	0,7625
673,75	0,7872	0,8209	0,8756	0,8985	0,9117	0,9301	0,9839	0,9977	1	0,9996	0,9558	0,9293	0,9295	0,9302	0,9333	0,9346	0,8638	0,8217	0,7630
681,25	0,7883	0,8218	0,8756	0,8984	0,9116	0,9307	0,9844	0,9973	0,9996	1	0,9583	0,9317	0,9315	0,9320	0,9352	0,9366	0,8651	0,8228	0,7638
708,75	0,7074	0,7332	0,7682	0,7853	0,8061	0,8715	0,9539	0,9686	0,9558	0,9583	1	0,9558	0,9461	0,9443	0,9492	0,9562	0,8787	0,8394	0,7883
753,75	0,7606	0,7809	0,8032	0,8012	0,8100	0,8285	0,8982	0,9299	0,9293	0,9317	0,9558	1	0,9954	0,9943	0,9969	0,9992	0,9679	0,9416	0,9030
761,25	0,7733	0,7952	0,8183	0,8167	0,8244	0,8374	0,8992	0,9282	0,9295	0,9315	0,9461	0,9954	1	0,9996	0,9986	0,9967	0,9737	0,9492	0,9161
764,375	0,7795	0,8012	0,8245	0,8226	0,8300	0,8419	0,9008	0,9286	0,9302	0,9320	0,9443	0,9943	0,9996	1	0,9991	0,9964	0,9755	0,9516	0,9186
767,5	0,7794	0,8005	0,8243	0,8226	0,8304	0,8439	0,9042	0,9323	0,9333	0,9352	0,9492	0,9969	0,9986	0,9991	1	0,9988	0,9747	0,9499	0,9140
778,75	0,7671	0,7882	0,8127	0,8120	0,8208	0,8388	0,9052	0,9349	0,9346	0,9366	0,9562	0,9992	0,9967	0,9964	0,9988	1	0,9702	0,9441	0,9052
865	0,7627	0,7790	0,7892	0,7804	0,7862	0,7957	0,8357	0,8615	0,8638	0,8651	0,8787	0,9679	0,9737	0,9755	0,9747	0,9702	1	0,9955	0,9771
885	0,7423	0,7561	0,7596	0,7480	0,7534	0,7645	0,7963	0,8200	0,8217	0,8228	0,8394	0,9416	0,9492	0,9516	0,9499	0,9441	0,9955	1	0,9908
900	0,7088	0,7190	0,7110	0,6923	0,6970	0,7111	0,7401	0,7625	0,7630	0,7638	0,7883	0,9030	0,9161	0,9186	0,9140	0,9052	0,9771	0,9908	1

Lisa 3. Peakomponentide varieeruvus

Tabel 6. Peakomponentide varieeruvuse kirjeldamine

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standardhälve	4,0964	1,1181	0,81327	0,40747	0,28043	0,2044	0,0964	0,06175
Osakaal koguvarieeruvusest	0,8832	0,0658	0,03481	0,00874	0,00414	0,0022	0,00049	0,0002
Kumulatiivne varieeruvus	0,8832	0,949	0,98379	0,99253	0,99667	0,9989	0,99936	0,99956

	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
Standardhälve	0,05873	0,04498	0,0356	0,02739	0,02371	0,01116	0,007452	0,006501
Osakaal koguvarieeruvusest	0,00018	0,00011	0,00007	0,00004	0,00003	0,00001	0	0
Kumulatiivne varieeruvus	0,99974	0,99985	0,99992	0,99995	0,99998	0,99999	0,99999	1

	PC17	PC18	PC19
Standardhälve	0,006265	0,005292	0,002639
Osakaal koguvarieeruvusest	0	0	0
Kumulatiivne varieeruvus	1	1	1

Lisa 4. Peakomponentide kaalud

Tabel 7. Peakomponentide kaalud lainepikkustel standardiseeritud peegeldumistegurite jaoks

Lainepikkused	PC1	PC2	PC3	PC4
400	0.21145	-0.30819	-0.37932	0.41382
412.5	0.2172	-0.31139	-0.32352	0.30413
442.5	0.22403	-0.32408	-0.18131	0.04068
490	0.2245	-0.32373	-0.04733	-0.25701
510	0.22607	-0.30618	0.00765	-0.33189
560	0.22729	-0.22818	0.16299	-0.43911
620	0.23359	-0.09108	0.30032	-0.06845
665	0.23555	-0.02237	0.30688	0.05012
673.75	0.23646	-0.04181	0.27549	0.04503
681.25	0.23672	-0.03982	0.27457	0.05562
708.75	0.22983	0.13158	0.31984	0.23005
753.75	0.23595	0.2073	0.02536	0.19184
761.25	0.23729	0.19317	-0.00946	0.11924
764.375	0.23784	0.18581	-0.01932	0.10103
767.5	0.23808	0.18456	-0.0065	0.11833
778.75	0.23717	0.19553	0.02415	0.15326
865	0.23033	0.24761	-0.20465	-0.15407
885	0.22362	0.27447	-0.27931	-0.26674
900	0.21309	0.31498	-0.36074	-0.3267

Lisa 5. Lõplik mudel

```
Coefficients :
                Estimate  Std. Error t-value  Pr(>|t|)
B:(intercept)  -167.01298 20560.11220 -0.0081  0.993519
C:(intercept)    -8.21694     2.01871 -4.0704 4.693e-05 ***
M:(intercept)    -1.22954     0.27337 -4.4977 6.870e-06 ***
V:(intercept)   -29.92103    14.75706 -2.0276  0.042604 *
B:PC1           -39.47463  4948.68562 -0.0080  0.993636
C:PC1            -2.14477     0.51901 -4.1324 3.590e-05 ***
M:PC1            -0.32889     0.12713 -2.5869  0.009684 **
V:PC1            -4.58064     2.38953 -1.9170  0.055242 .
B:PC2            77.54677  9193.31222  0.0084  0.993270
C:PC2            -4.05625     0.92730 -4.3742 1.219e-05 ***
M:PC2            -0.39597     0.36775 -1.0768  0.281589
V:PC2            22.28424    10.76090  2.0709  0.038373 *
B:PC3           -20.81286  6434.02332 -0.0032  0.997419
C:PC3            -6.99232     1.64192 -4.2586 2.057e-05 ***
M:PC3            -0.84905     0.39110 -2.1709  0.029937 *
V:PC3             5.05846     3.14238  1.6098  0.107451
B:PC4            97.81560 14229.53092  0.0069  0.994515
C:PC4            -4.45923     1.93684 -2.3023  0.021317 *
M:PC4            -0.38753     0.65492 -0.5917  0.554034
V:PC4            37.21829    18.83679  1.9758  0.048174 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Joonis 5. R-i väljund nelja peakomponendiga mudeli parameetrite kohta

Joonisel 5 toodud R-i väljundis olevad tähised:

- B – *Pruun*;
- C – *Selge*;
- M – *Mõõdukas*;
- V – *Väga sogane*;
- Baas T – *Sogane*.

Lisa 6. Mudeliga prognoositud veetüüpide võrdlus tegelikkusega

Tabel 8. Viie peakomponendiga mudeli prognooside võrdlus tegelikega

tegelik / mudel	*Pruun*	*Selge*	*Mõõdukas*	*Sogane*	*Väga sogane*
Pruun	9	0	0	0	0
Selge	0	31	3	2	0
Mõõdukas	0	0	0	2	0
Sogane	0	3	25	90	1
Väga sogane	0	0	1	0	13

Tabel 9. Kuue peakomponendiga mudeli prognooside võrdlus tegelikega

tegelik / mudel	*Pruun*	*Selge*	*Mõõdukas*	*Sogane*	*Väga sogane*
Pruun	9	0	0	0	0
Selge	0	32	1	1	0
Mõõdukas	0	0	3	4	0
Sogane	0	2	25	89	0
Väga sogane	0	0	0	0	14

Tabel 10. Seitsme peakomponendiga mudeli prognooside võrdlus tegelikega

tegelik / mudel	*Pruun*	*Selge*	*Mõõdukas*	*Sogane*	*Väga sogane*
Pruun	9	0	0	0	0
Selge	0	33	0	1	0
Mõõdukas	0	1	8	5	0
Sogane	0	0	21	88	0
Väga sogane	0	0	0	0	14

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Getter Põru (sünnikuupäev 17.10.1995),

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Veekogude klassifitseerimine satelliidiandmetelt multinomiaalse logistilise mudeliga“, mille juhendajad on Ene Käärik ja Kristi Uudeberg,

   (a) reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace'i lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

   (b) üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartu, 08.05.2018